SCAI-QReCC Shared Task on Conversational Question Answering



Svitlana Vakulenko
University of Amsterdam (Amazon)

Johannes Kiesel
Bauhaus-Universität Weimar

Maik Fröbe
Martin Luther University Halle-Wittenberg

SCAI: Search-Oriented Conversational AI

https://scai.info

SCAI: Search-Oriented Conversational AI

https://scai.info

- Conversational Search - the task of retrieving relevant information using a conversational interface / automating an **information-seeking dialogue**

Svitlana Vakulenko. Knowledge-based Conversational Search. PhD thesis. Technical University of Vienna.
Avishek Anand, Lawrence Cavedon, Matthias Hagen, Hideo Joho, Mark Sanderson, Benno Stein: Conversational Search - A Report from Dagstuhl Seminar 19461. CoRR abs/2005.08658 (2020)

SCAI: Search-Oriented Conversational AI

https://scai.info

- Conversational Search - the task of retrieving relevant information using a conversational interface / automating an **information-seeking dialogue**

2017 ICTIR

2018 EMNLP

2019 WebConf

2019 IJCAI

2020 EMNLP

2021 online event

2022 SIGIR

SCAI: Search-Oriented Conversational AI

https://scai.info

- Question Answering
- User Satisfaction & Dialogue Breakdown
- Dialogue Personalization
- Mixed Initiative
- **SCAI-QReCC Shared Task on Conversational QA**

Conversational QA Example

Anaphora

Ellipsis

Question: Tell me about the benefits of **Yoga**?
Answer: Increased flexibility, muscle strength…
URL: https://osteopathic.org/what-is-osteopathic-medicine/benefits-of-yoga

Question: Does **it** help in reducing stress?
Rewrite: Does **Yoga** help in reducing stress?
Answer: Yoga may help reduce stress, lower blood pressure, and lower your heart rate.
URL: https://www.mayoclinic.org/healthy-lifestyle/stress-management/in-depth/yoga/art-20044733

Question: What are some of the main types?
Rewrite: What are some of the main types **of Yoga**?
Answer: Hatha, Kundalini, Ashtanga, …
URL: https://www.mindbodygreen.com/articles/the-11-major-types-of-yoga-explained-simply

Question: What are common poses in Kundalini Yoga?
Rewrite: What are common poses in Kundalini Yoga?
Answer: Lotus Pose, Celibate Pose, Perfect Pose, …
URL: https://www.kundaliniyoga.org/Asanas

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, Srinivas Chappidi: Open-Domain Question Answering Goes Conversational via Question Rewriting. NAACL-HLT 2021: 520-534

QReCC

- 14K English QA dialogues
- 81K question-answer pairs
- 54M passages from web pages

max Qs/dialogue	12
avg Qs/dialogue	6
min Qs/dialogue	5

https://github.com/apple/ml-qrecc

https://zenodo.org/record/5543685#.YV_OEC0RppR

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, Srinivas Chappidi: Open-Domain Question Answering Goes Conversational via Question Rewriting. NAACL-HLT 2021: 520-534

Conversational QA Subtasks

1. Question Rewriting
2. Passage Retrieval
3. Question Answering

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, Srinivas Chappidi: Open-Domain Question Answering Goes Conversational via Question Rewriting. NAACL-HLT 2021: 520-534

Conversational QA Evaluation

1. Question Rewriting: ROUGE-1 Recall
2. Passage Retrieval: Mean Reciprocal Rank
3. Question Answering: F-measure

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, Srinivas Chappidi: Open-Domain Question Answering Goes Conversational via Question Rewriting. NAACL-HLT 2021: 520-534

Research Questions

- RQ1: Are there alternative correct answers?
- RQ2: What is the best approach for conversational QA?
- RQ3: How to evaluate this task?

Results

- 30 runs
  - 4 teams: Rachael, Rali, Torch, Ultron
  - 3 baselines: gpt3, simple, basic
- 16,736 answers

https://en.wikipedia.org/wiki/List_of_fictional_robots_and_androids

Metrics

- **BERT** (BERTScore, Zhang et al., 2020)
- **POSS** (POSSCORE, Liu et al., 2021)
- **SAS** (Semantic Answer Similarity, Risch et al., 2021)
- **BKPQA & RKPQA** (BERTScore & ROUGE-L KPQA, Lee et al., 2021)

Baselines

- simple
  - BM25 + question overlap heuristic
- basic
  - Answer = Question
- gpt3
  - 50 USD via OpenAI API

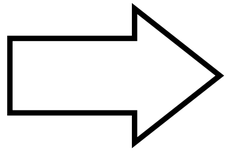
Results

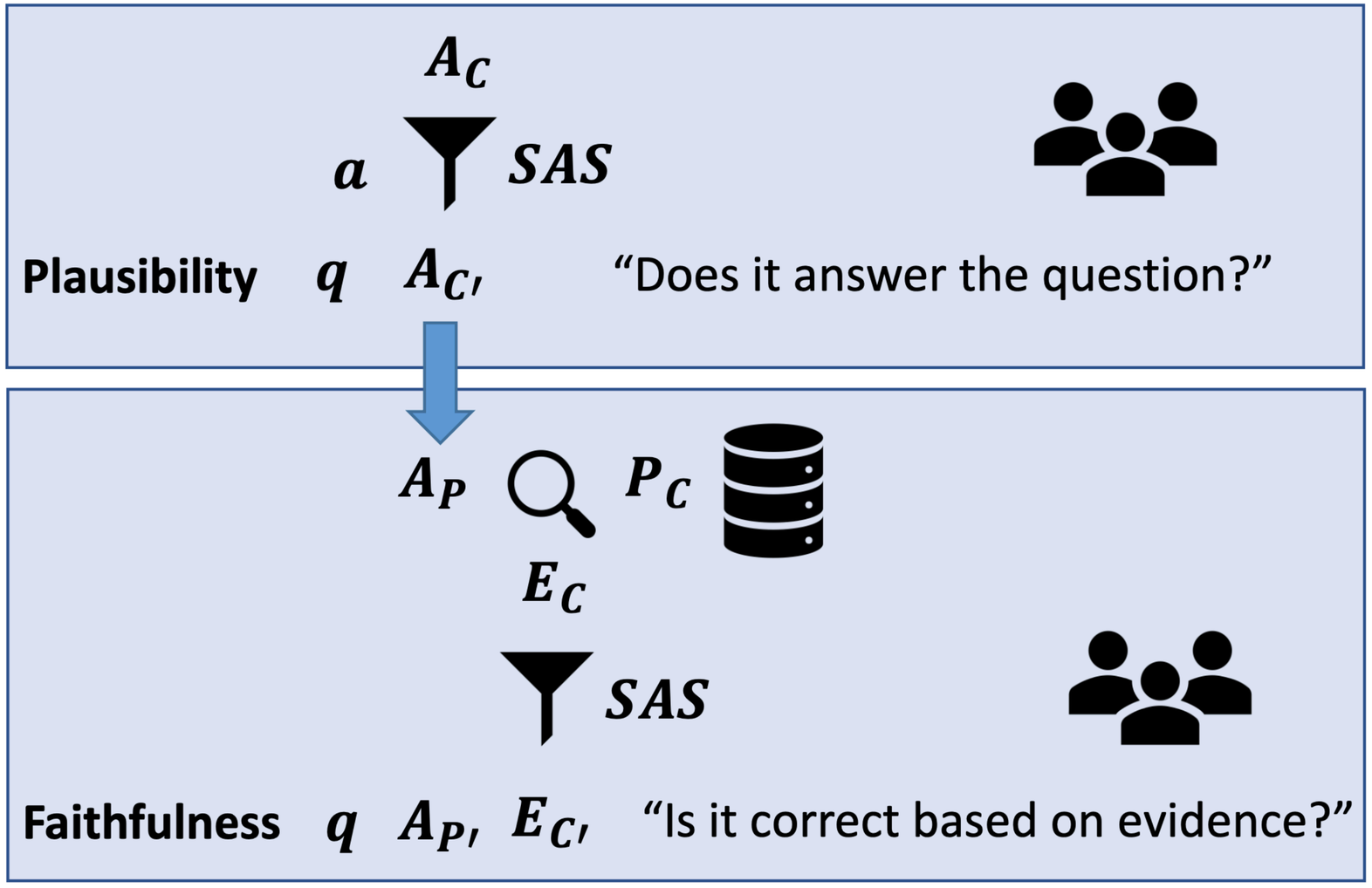
Team	Run	QR	MRR	EM	F1	R1	POSS	SAS	BERT	BKPQA	RKPQA
Original questions											
-	Basic baseline	-	-	0.000	0.114	0.095	1.283	0.207	0.422	0.432	0.064
-	GPT3 baseline	-	-	0.001	0.149	0.148	1.305	0.264	0.448	0.467	0.134
-	Simple baseline	0.571	0.065	0.001	0.067	0.150	1.490	0.162	0.367	0.426	0.097
rachael	2021-09-04-10-38-07	-	0.056	0.002	0.138	0.193	**1.583**	0.163	0.410	0.476	0.135
rachael	2021-09-08-07-07-57	0.675	0.135	0.006	0.187	**0.226**	1.570	**0.277**	**0.452**	**0.498**	0.175
rachael	2021-09-08-07-09-57	0.682	0.128	0.006	0.186	0.226	1.558	0.269	0.448	0.494	0.175
rachael	2021-09-08-15-40-34	0.679	0.133	0.007	0.176	0.211	1.456	0.254	0.420	0.460	0.164
rachael	2021-09-08-21-49-44	0.681	0.130	0.008	0.177	0.211	1.461	0.246	0.422	0.462	0.167
rachael	2021-09-15-09-05-06	0.673	**0.158**	**0.011**	0.179	0.212	1.333	0.254	0.405	0.444	0.172
rachael	2021-09-15-09-06-44	0.681	0.150	0.010	0.179	0.211	1.369	0.249	0.408	0.449	0.169
rachael	2021-09-15-09-07-49	0.676	0.157	0.010	0.187	0.219	1.399	0.264	0.418	0.457	0.175
rachael	2021-09-15-09-08-40	**0.685**	0.149	0.010	**0.189**	0.222	1.458	0.259	0.428	0.470	**0.178**
torch	usi_T5_raw2	0.657	0.082	0.001	0.137	0.200	1.451	0.221	0.415	0.467	0.117

Manual Annotations

Question q

Answers A





SCAI-QReCC

QReCC

Conversational Question: "Do **NPs** or **PAs** make more?"

Rewritten Question: "Do **nurse practitioners** or **physician assistants** make more?"

Human Answer: "According to the Bureau of Labor and Statistics (BLS), the **2017** median pay for **Nurse Practitioners** is $110,930 per year, or $53.33 per hour. In comparison, 2017 median pay for **Physician Assistants** is $104,860 per year, or $50.41 per hour."

Submitted

Generated Answer: "The average salary for a **nurse practitioner** is $65,000. The average salary for a **physician's assistant** is $30,000."

SAS

Extracted Answer: "The BLS reports that the median annual wage for **nurse practitioners** was $109,820 as of May **2019** , while the median annual wage for **physician assistants** reached $112,260 during the same month."

Annotation Guideline

Submitted

SAS

Read the question and the answer carefully

Question: What is Ohio Valley Wrestling?

Answer: Ohio Valley Wrestling (OVW) is a professional wrestling association based in Des Moines, Iowa.

Can you tell that the answer to the question above is correct based on the information provided in the text below?

Text: ovwrestling .com Ohio Valley Wrestling (OVW) is an American independent professional wrestling promotion based in Louisville, Kentucky .

- ○ The text is relevant to the question and it contains information sufficient to judge the answer as correct
- ○ The text is relevant to the question, but it does not contain information to judge the answer as correct
- ○ The text is irrelevant to the question

Team	Run	Question	Plausible	Implausible	Malformed	Faithful	Unfaithful
rachael	2021-09-04-10-39-42	rewritten	**183**	5	4	**37**	2
rachael	2021-09-08-21-49-44	original	133	6	4	30	1
rachael	2021-09-08-07-07-57	original	120	4	5	30	0
rachael	2021-09-15-09-07-49	original	103	4	6	29	1
-	GPT3 baseline	original	149	4	8	28	3
ultron	rag-bm25_100	rewritten	173	15	6	27	2
rachael	2021-09-06-09-21-43	rewritten	158	4	3	26	**4**
ultron	2021-09-08-15-04-28	rewritten	149	**16**	6	24	1
rachael	2021-09-15-19-36-31	rewritten	132	2	2	24	0
rachael	2021-09-15-09-06-44	original	73	0	4	22	1
rachael	2021-09-08-07-09-57	original	75	2	4	16	1
rali-qa	2021-09-09-13-01-07	rewritten	33	6	11	16	1
rachael	2021-09-08-15-40-34	original	41	6	2	14	3
torch	usi T5 raw2	original	36	7	**16**	14	0
ultron	2021-09-04-17-28-07	rewritten	117	13	7	13	0
rachael	2021-09-15-09-08-40	original	52	4	4	10	0
ultron	BART-large-top1BM25	rewritten	29	3	11	10	0
rachael	2021-09-15-09-05-06	original	52	2	1	9	0
rachael	2021-09-04-10-38-07	original	41	2	0	6	1
-	Simple baseline	rewritten	14	2	3	1	0
-	Simple baseline	original	0	0	1	0	0
Total			1863	107	108	386	21

Results

Team	Run	QR	MRR	EM	F1	R1	POSS	SAS	BERT	BKPQA	RKPQA
Original questions											
-	Basic baseline	-	-	0.000	0.114	0.095	1.283				
-	GPT3 baseline	-	-	0.001	0.149	0.148	1.305				
-	Simple baseline	0.571	0.065	0.001	0.067	0.150	1.490				
rachael	2021-09-04-10-38-07	-	0.056	0.002	0.138	0.193	**1.583**				
rachael	2021-09-08-07-07-57	0.675	0.135	0.006	0.187	**0.226**	1.570				
rachael	2021-09-08-07-09-57	0.682	0.128	0.006	0.186	0.226	1.558				
rachael	2021-09-08-15-40-34	0.679	0.133	0.007	0.176	0.211	1.456				
rachael	2021-09-08-21-49-44	0.681	0.130	0.008	0.177	0.211	1.461				
rachael	2021-09-15-09-05-06	0.673	**0.158**	**0.011**	0.179	0.212	1.333				
rachael	2021-09-15-09-06-44	0.681	0.150	0.010	0.179	0.211	1.369				
rachael	2021-09-15-09-07-49	0.676	0.157	0.010	0.187	0.219	1.399				
rachael	2021-09-15-09-08-40	**0.685**	0.149	0.010	**0.189**	0.222	1.458				
torch	usi_T5_raw2	0.657	0.082	0.001	0.137	0.200	1.451				
Human rewritten questions											
-	Basic baseline		-	0.000	0.224	0.205	1.555				
-	Simple baseline		**0.398**	0.001	0.098	0.282	1.666				
rachael	2021-09-04-10-39-42		0.359	0.011	0.267	0.331	**1.674**				
rachael	2021-09-06-09-21-43		0.359	0.018	0.290	0.339	1.649				
rachael	2021-09-15-19-36-31		0.385	**0.028**	**0.302**	**0.345**	1.618				
rali-qa	2021-09-09-13-01-07		0.269	0.003	0.166	0.212	1.385				
ultron	2021-09-04-17-28-07		-	0.001	0.183	0.186	1.357				
ultron	2021-09-08-15-04-28		-	0.015	0.261	0.258	1.565				
ultron	2021-09-08-15-07-30		-	0.001	0.187	0.189	1.380				
ultron	2021-09-08-15-08-00		-	0.004	0.247	0.236	1.597	0.379	0.536	0.525	0.177
ultron	bart-large_top1bm25		-	0.000	0.017	0.017	0.150	0.111	0.046	0.048	0.016
ultron	distilbart-xsum-12-1_top1bm25		-	0.000	0.019	0.020	0.170	0.113	0.050	0.054	0.018
ultron	distilbart-xsum-12-3_top1bm25		-	0.000	0.022	0.023	0.175	0.117	0.052	0.056	0.021

Team	Run	Question	Plausible	Implausible	Malformed	Faithful	Unfaithful
rachael	2021-09-04-10-39-42	rewritten	**183**	5	4	**37**	2
rachael	2021-09-08-21-49-44	original	133	6	4	30	1
rachael	2021-09-08-07-07-57	original	120	4	5	30	0
rachael	2021-09-15-09-07-49	original	103	4	6	29	1
-	GPT3 baseline	original	149	4	8	28	3
ultron	rag-bm25_100	rewritten	173	15	6	27	2
rachael	2021-09-06-09-21-43	rewritten	158	4	3	26	**4**
ultron	2021-09-08-15-04-28	rewritten	149	**16**	6	24	1
rachael	2021-09-15-19-36-31	rewritten	132	2	2	24	0
rachael	2021-09-15-09-06-44	original	73	0	4	22	1
rachael	2021-09-08-07-09-57	original	75	2	4	16	1
rali-qa	2021-09-09-13-01-07	rewritten	33	6	11	16	1
rachael	2021-09-08-15-40-34	original	41	6	2	14	3
torch	usi T5 raw2	original	36	7	**16**	14	0
ultron	2021-09-04-17-28-07	rewritten	117	13	7	13	0
rachael	2021-09-15-09-08-40	original	52	4	4	10	0
ultron	BART-large-top1BM25	rewritten	29	3	11	10	0
rachael	2021-09-15-09-05-06	original	52	2	1	9	0
rachael	2021-09-04-10-38-07	original	41	2	0	6	1
-	Simple baseline	rewritten	14	2	3	1	0
-	Simple baseline	original	0	0	1	0	0
Total			1863	107	108	386	21

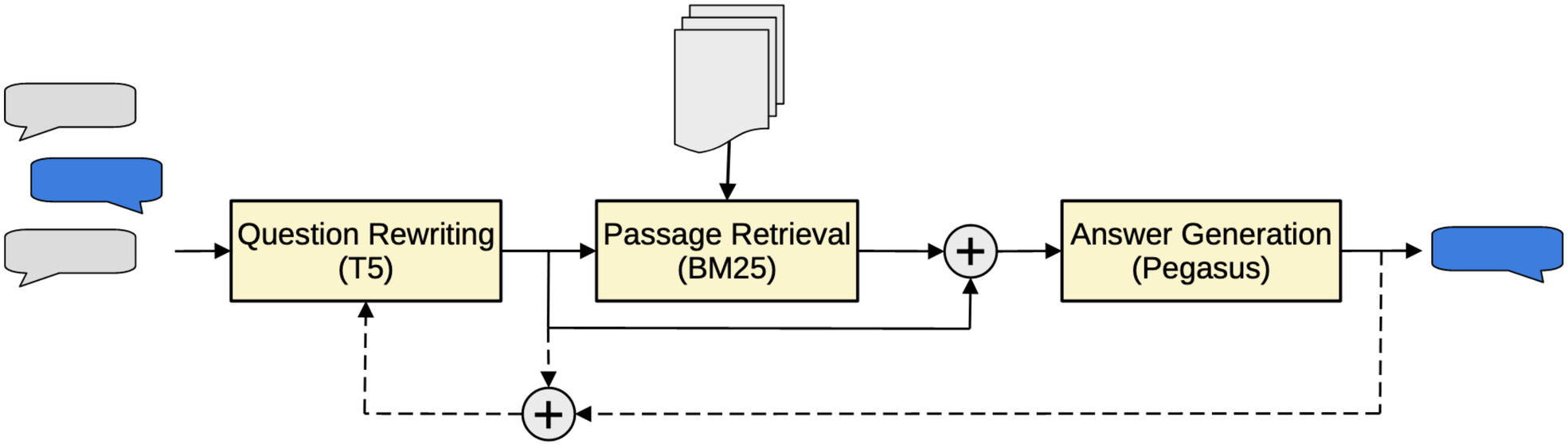
Conclusion

RQ1: Are there alternative correct answers?

- Public dataset: https://doi.org/10.5281/zenodo.5749472
  - + **answer spaces**: multiple correct answers

Conclusion

RQ2: What is the best approach for conversational QA?



Gonçalo Raposo, Rui Ribeiro, Bruno Martins, Luísa Coheur: Question Rewriting? Assessing Its Importance for Conversational Question Answering. ECIR (2) 2022: 199-206

Conclusion

RQ3: How to evaluate Conversational QA?

- Evolution of the **evaluation guidelines**
  - Provide **evidence** for the generated answers
  - Evaluate answers with **ROUGE-1 Recall**

Future Work

- 2022 Evaluation with multiple correct answers
- 2023 Human-in-the-loop evaluation

SCAI: Search-Oriented Conversational AI

Workshop at SIGIR 2022

July 15, Madrid/Online

This workshop is intended as a **discussion platform on Conversational AI for intelligent information access**, bringing together researchers and practitioners across natural language processing, information retrieval, machine learning and human-computer interaction fields. Among other topics, we will discuss design, evaluation and human factors in relation to automating information-seeking dialogues. The workshop will also feature a shared task on Conversational Question Answering.

SCAI 2022 is a 7th iteration of SCAI. This year, the workshop will focus on the discussion between researchers from different fields. It is therefore organized as a **strictly non-archival venue**, as an opportunity to present papers accepted to other venues in an interdisciplinary meeting specifically focused on search-based conversational AI. SCAI 2022 is a **hybrid workshop** at SIGIR 2022, taking place on 15 July (full day).

Check out the recordings of SCAI'21 on our YouTube channel

https://sigir.org/sigir2022/